Set-valued Bayesian Inference with Probabilistic Equivalence

Hoel Le Capitaine
LINA (UMR CNRS 6241), École Polytechnique de Nantes
Rue C. Pauc, La Chantrerie, 44306 Nantes, France.
hoel.lecapitaine@univ-nantes.fr

Abstract

In this paper, a unified view of the problem of class-selection with Bayesian classifiers is presented. Selecting a subset of classes instead of singleton allows 1) to reduce the error rate and 2) to propose a reduced set to another classifier or an expert. This second step provides additional information, and therefore increases the quality of the result. The proposed framework, based on the evaluation of the probabilistic equivalence, allows to retrieve the class-selective frameworks that have been proposed in the literature. Several experiments show the effectiveness of this generic proposition.

1. Introduction and basic material

The process of accurately recognizing, or discriminating, objects in databases is a fundamental task in data analysis. Considered as a pattern recognition problem, there have been many propositions for the classification of the objects. Naturally, the more a priori information is known, the more the classification algorithm can be built according to this knowledge, therefore leading to a powerful recognition system. In the special case where a priori probabilities and conditional densities are known, the Bayes decision rule is known to be optimal by minimizing the error rate [12]. However, real distributions are never known in advance, so that the models do not reflect the data. Moreover, in many recognition problems, the data that must be classified is issued from mixed and/or noisy classes. In particular, it is not uncommon to find classes that are overlapping in the feature space, or some samples that do not belong to any class of the learning set.

Number of investigation in the field of pattern recognition focus on problems with a large number of classes (*e.g.* face identification, image classification, character recognition and so on ...). Therefore, the possibility of selecting a small subset of classes, which can be associated to a new, larger class, containing the previous classes, shows a growing interest [2]. Finally, another growing interest resides in multi-label classification [13], where a sample can be associated to a subset of true labels.

In this paper, the Bayesian statistical decision theory is taken as the basis of the analysis. Therefore, we start with the following three factors: the distribution family $p(\mathbf{x}|\theta)$, prior distribution for the parameters $\pi(\theta)$ and a loss function $\ell(\theta, \alpha)$, where α is an action of the decision space $\mathcal{A}$ [12]. An effective comparison criterion is the a posteriori expected loss, which can be written as

$$R(\alpha, \mathbf{x}) = \int_{\Theta} \ell(\theta, \alpha) p(\theta|\mathbf{x}) \mathrm{d}\theta \tag{1}$$

where Θ is the state space (we consider in the sequel the discrete case where the cardinal of Θ is equal to c). The posterior probability $p(\theta|\mathbf{x})$ is obtained thanks to the Bayes theorem, knowing the distribution family and priors on $\theta \in \Theta$. Under the Bayes principle, the optimal rule is obtained by choosing for $\mathbf{x}$ the action α that minimize the expected loss:

$$\alpha(\mathbf{x}) = \arg\min_{\alpha \in \mathcal{A}} R(\alpha, \mathbf{x})$$

Naturally, if the expected loss (1) is minimum for all $\mathbf{x}$, then the overall risk is also minimized. If one seeks to minimize the error probability of classification, then the zero-one loss function is used. One may also allow other actions than a binary and strict association to classes. For instance, the reject option consists in adding another action in $\mathcal{A}$ [1, 14]. The action leads to refuse, or withhold, the decision for the current sample. This is particularly useful in close cases i.e. when the largest posterior probabilities are close. Naturally, the *no decision* action must have a cost, or a loss, that needs to be modeled under the Bayes minimum risk setting.

In this paper, we are interested in an even more increased action space $\mathcal{A}$. In particular, we consider the

power set of Θ. Therefore, each sample $\mathbf{x}$ can be associated to one element of the power set. The subset selection procedure is described in the next section.

2. Subset selection

The basic principle of set-valued classification is allowing to select a subset of classes of interest. The subset can subsequently be used as whether an entry for another classification with large error costs, or an output for multi-label classification. Depending on the application, the loss function can be designed differently. We focus in this paper on the first one. In this first application, the goal is to reduce the error probability by selecting more than one class, but selecting all classes does not provide any profit of the algorithm. Therefore, a compromise between the number of selected classes and the error probability must be found. A simple solution consists in dividing the loss function into two parts, the error loss ℓ_e and the selection loss ℓ_n [5]. The error loss can be adapted from the so-called symmetrical loss function

$$\ell_e(\theta_i, \alpha_j) = \begin{cases} 0 & \text{if } \theta_i \in \mathcal{A}_j \\ C_e & \text{otherwise} \end{cases}$$

where $\mathcal{A}_j$ is the selected subset of $\mathcal{A}$ with action α_j, and C_e the cost of an error. The selection loss is a function of the number of selected classes

$$\ell_n(\alpha_j) = C_n|\mathcal{A}_j|$$

where C_n is the cost of selecting classes, and $|.|$ is the cardinal. From now on, we suppose that posterior probabilities are sorted such that $p(\theta_{k+1}|\mathbf{x}) \leq p(\theta_k|\mathbf{x})$ for all k in $\{1, \cdots, c-1\}$. With this formulation, one can prove that the following decision rule $\alpha_j(\mathbf{x})$ yields an optimum trade-off between the error and the number of selected classes [5]

$$n^\star(\mathbf{x}) = \min_{k \in [1,c]} \{k : p(\theta_{k+1}|\mathbf{x}) \leq t\},\ p(\theta_{c+1}|\mathbf{x}) \equiv 0 \tag{2}$$

where $|\mathcal{A}_j| = n^\star(\mathbf{x})$, $\mathcal{A}_j$ is composed of the $n^\star(\mathbf{x})$ largest posterior probabilities, and t a threshold in the unit interval defined by C_n/C_e. Note that the reverse, i.e. subset rejection, has been proposed in [9]. The threshold t is defined by ratio of costs so that it is mostly application dependent, but a generic evaluation can be proposed, as it is proposed in Section 4

Another proposition, coming from Horiuchi in [6], can be used to define a set-valued output of classes. The corresponding decision rule is defined by

$$n^\star(\mathbf{x}) = \min_{k \in [1,c]} \{k : 1 - \big(p(\theta_k|\mathbf{x}) - p(\theta_{k+1}|\mathbf{x})\big) \leq t\}, \tag{3}$$

using the same convention $p(\theta_{c+1}|\mathbf{x}) \equiv 0$. However, this proposition is not obtained using a loss function and can be seen as an heuristic.

Another heuristic, proposed in [8], defined by

$$n^\star(\mathbf{x}) = \min_{k \in [1,c]} \{k : \frac{p(\theta_{k+1}|\mathbf{x})}{p(\theta_k|\mathbf{x})} \leq t\}, \tag{4}$$

also uses a notion of similarity between consecutive posterior probabilities. We restrict in this paper to standard approaches of class-selection, but it should be noted that other strategies based on blockwise similarities [10] or support vector machines [4] have been proposed.

3. Probabilistic equivalence

In this section, we propose to design a new decision rule based on the equivalence of posterior probabilities. The equivalence is obtained by considering a probabilistic metric (PM) space where a convenient metric is chosen between two values. Formally, a metric space consists of a set X and a metric d allowing to compute distances between two points u, v lying in X. A PM space replaces the distance $d(u, v)$ between the two points by considering a distribution function F_{uv}, whose value $F_{uv}(x)$, for any x in X, corresponds to the probability that $d(u, v) \leq x$. However, one of the most important property of distances is that they hold the triangle inequality $d(u, w) \leq d(u, v) + d(v, w)$, for $(u, v, w) \in X^3$. The corresponding problem with distribution function relies on the comparison and relationships of F_{uw}, F_{uv} and F_{vw}. This is the rationale under the proposition of Menger [11], introducing the following inequality:

$$F_{uw}(x + y) \geq T(F_{uv}(x), F_{vw}(y))$$

where T is a triangular norm (t-norm), *i.e.* a commutative, associative and monotone binary function, having 1 as identity, see [7] for details. Let us consider the t-norm defined by

$$T(x, y) = \big(\max\big(x^\lambda + y^\lambda - 1, 0\big)\big)^{1/\lambda}$$

where $(x, y) \in [0, 1]^2$ and $\lambda \in [-\infty, \infty]$. It leads to the following residual implication between x and y

$$I(x, y) = \begin{cases} \big(1 + y^\lambda - x^\lambda\big)^{1/\lambda} & \text{if } x \geq y \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

Based on this implication, a T-equivalence is obtained with

$$E(x, y) = \min\left(I(x, y), I(y, x)\right) \tag{6}$$

Applied to posterior probabilities, we propose to define the new decision rule as

$$n^\star(\mathbf{x}) = \min_{k\in[1,c]} \{k : E\big(p(\theta_k|\mathbf{x}), p(\theta_{k+1}|\mathbf{x})\big) \leq t\}, \tag{7}$$

called here after PE for Probabilistic Equivalence. Now, let us consider some particular cases of (7) when using the equivalence (6) obtained by the implication defined by (5).

- if λ is set to $-\infty$, then $E(x, y) = y$, giving the decision rule (2).
- if λ is set to 0, then $E(x, y)$ reduces to y/x, giving the decision rule (4).
- finally, if $\lambda = 1$, then $E(x, y) = 1 + y - x$, which gives the decision rule (3).

Due to lack of place, the proofs of these statements will be given in a longer paper. As can be seen, the proposed rule allows to retrieve the set-valued decision rules of the literature.

4. Experiments

In theory, the rule defined by (2) is the optimum decision rule in the sense that there are no other rules yielding a lower error rate for a given average number of selected classes. This optimum is reached when the distribution of the data is known and true which is rarely the case in practice. Moreover, this is an optimum rule with respect to the average number of selected classes, which may not be the only criteria that must be considered. Therefore, in this section a comparative study between the three common rules and the new one is proposed.

4.1. Experimental setup and evaluation metric

In the experiments, 4 datasets publicly available [3] are used: Vowel, Letter, Segment and USPS. Important statistics of the datasets are given in Table 1. These datasets are chosen for two particular reasons. First, they present a large variety in terms of number of classes, number of features and number of samples. Second they are all coming from applications where allowing to select a subset of classes is very interesting: character recognition and image segmentation. In order to select the parameter λ, we use a grid search in the range $[-10, 10]$ based on a 2-fold cross validation that is repeated 5 times. Two different classifiers are used in this study: a naive Bayes (NB) classifier and a quadratic classifier (QB) assuming normal densities. It is important to note that what is evaluated here is the selection step, and not the classifiers. The two classifiers are used in order to assess the consistency of a possible superiority of a decision rule for a given classifier.

Table 1. Datasets used in the experiments.

Dataset	#training	#testing	#classes	#feat.
Segment	2310	0	7	19
Vowel	528	462	11	10
Letter	15000	5000	26	16
USPS	7291	2007	10	256

Reject options, and more generally class-selective decision rules cannot be evaluated by considering only their corresponding accuracy. This is due to two major reasons. The first is that they generally use a specified threshold, therefore giving different classification and rejection rates. The second reason comes from the tradeoff they imply. For the reject option, the tradeoff to find is between a low error rate and a low rejection rate. For class-selective decision, the tradeoff concerns a low error rate, and a low average number of selected classes. Therefore, a common quality measure is to evaluate the area under the curve (Error(Rate) for the reject options, Error(Average) for class-selective). However, this evaluation measure is not adapted for the comparison of decision rules used on different classifiers, because each classifier provides different accuracies without rejecting samples. We propose to define the normalized area under the curve in order to overcome this problem. This evaluation measure must take into account the baseline performance of classifier, so that we define the normalized area under the curve ($nAUC$) by

$$nAUC = \frac{\int_0^1 (C(t) - C_0)\,\mathrm{d}t}{\int_0^1 (1 - C_0)\,\mathrm{d}t} \tag{8}$$

where C_0 is the baseline accuracy of the classifier (i.e. the classification rate without reject option), and t is the threshold used in the decision rule (7). The term $C(t)$ is the classification rate obtained by selecting subset of classes using (7). The convention is to say that the classification is good if one label of the subset is the true label of the sample.

It can be proved that $C_0 \leq C(t) \leq 1$ for any t in the unit interval (we have in particular $C(0) = 1$ and $C(1) = C_0$), so that $0 \leq nAUC \leq 1$. A $nAUC$ equal to 1 means that far any t, the error rate with subset selection is equal to zero, while $nAUC$ value equal to zero means that adding subset of classes does not increase the classification rate at all. Therefore, the higher the better for $nAUC$.

4.2. Results

The evaluation criteria (8) is computed for each dataset, each classifier and each decision rule. Results

Table 2. Normalized AUC for all datasets and all decision rules. Right column indicates the average rank of decision rules over all datasets.

		Datasets				
Classifier	Rule	Segment	Vowel	Letter	USPS	Avg. Rank
NB	Ha	63.61%	84.77%	81.61%	72.79%	3
	Horiuchi	64.80%	65.04%	70.59%	77.85%	3
	Le Capitaine	64.67%	82.17%	79.89%	78.13%	2.75
	PE	**72.38%**	**85.09%**	**84.94%**	**83.68%**	**1**
QB	Ha	92.50%	59.25%	83.44%	73.54%	3.5
	Horiuchi	89.04%	53.27%	85.09%	80.17%	3.25
	Le Capitaine	92.88%	60.33%	90.79%	79.50%	2.25
	PE	**94.07%**	**60.46%**	**91.07%**	**84.61%**	**1**

are given in Table 2. As can be seen in the table, the PE decision rule gives the best results for both classifiers and all datasets. Naturally, other comparisons using many other datasets would be required to assess a definitive superiority of the decision rule PE, but it gives encouraging results. Looking more in depth the results, one can say that one can observe a larger difference between rules for NB than for QB, which is explained by the quality of estimation of posterior probabilities. The $nAUC$ score is generally better for QB than for NB, due to the performances rates of individual methods. While the ratio rule (Le Capitaine) ranks second for both classifiers, the performances of Ha and Horiuchi rules are comparable. Some datasets lead to remarkably bad results, for instance Vowel for Horiuchi's rule, or USPS for Ha's rule. Finally, one can see that there is great benefit of introducing class selection for the USPS and Segment datasets.

5. Conclusion

In this paper, a generalized approach to class-selection is presented. Given a classifier providing posterior probabilities outputs, the proposed rule allows to retrieve the three class-selective decision rules proposed in the literature. In this paper, a simple grid search is used to find the parameter, but investigations on automatic learning of the parameter λ based on the correct set of classification of each sample is under study. As a potential future work, let us mention the multi-label multi-class classification, where each sample may belong to several classes.

References

[1] C. Chow. On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[2] J. del Coz, J. Diez, and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10:2273–2293, 2009.

[3] A. Frank and A. Asuncion. UCI machine learning repository, 2010. University of California, Irvine, School of Information and Computer Sciences.

[4] E. Grall-Maes and P. Beauseroy. Optimal decision rule with class-selective rejection and performance constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2073–2082, 2009.

[5] T. Ha. The optimum class-selective rejection rules. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.

[6] T. Horiuchi. Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, 31(10):1579–1588, 1998.

[7] E. P. Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005.

[8] H. Le Capitaine. *Aggregation operators for similarity measures. Application to ambiguity in pattern recognition*. PhD thesis, Univ. of La Rochelle, 2009.

[9] H. Le Capitaine and C. Frélicot. An optimum class-rejective decision rule and its evaluation. In *20th International Conference on Pattern Recognition, 2010*, pages 3312–3315, Istanbul, Turkey, 2010.

[10] H. Le Capitaine and C. Frélicot. A family of measures for best top-n class-selective decision rules. *Pattern Recognition*, 45(1):552–562, 2012.

[11] K. Menger. Statistical metrics. *Proc. National Academy of Science USA*, 28(12):535–537, 1942.

[12] C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag, 2007.

[13] G. Tsoumakas and I. Katakis. Multi-label classification. *International Journal of Data Warehousing & Mining*, 3(3):1–13, 2007.

[14] M. Yuan and B. Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11:111–130, 2010.